Introduction to NLP for Software Developers

Maciej Szymczak

AI Day Wroclaw
Volvo Group IT in Poland
29th November 2018

VOLVO
VOLVO GROUP

Agenda

- [x] What is NLP
- [x] How to do NLP
- [x] Some advanced topics
- [x] DEMO
- [x] Conclusion and Q&A

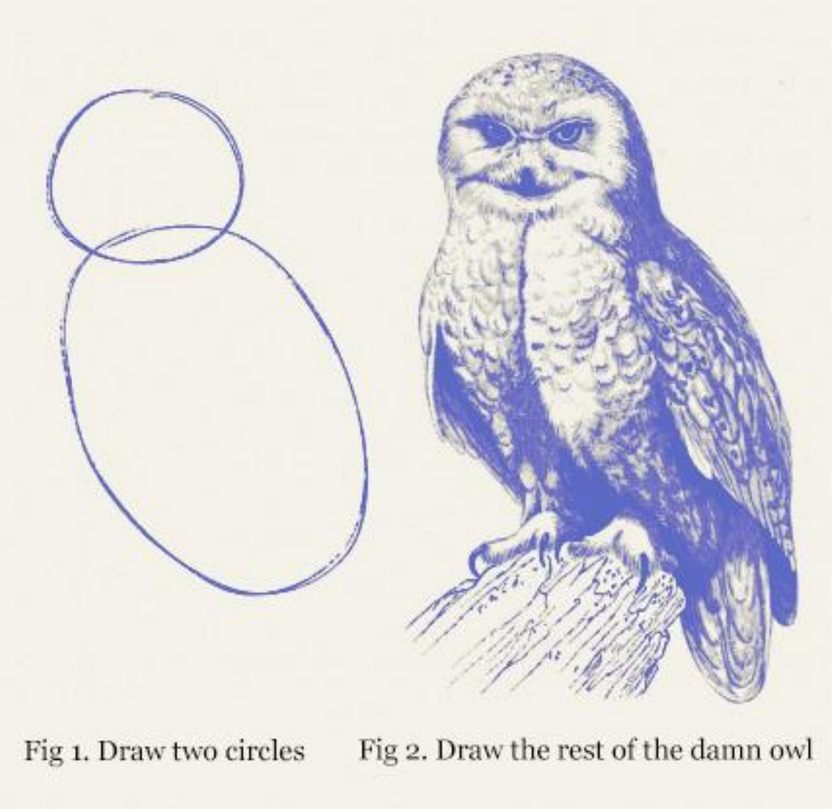
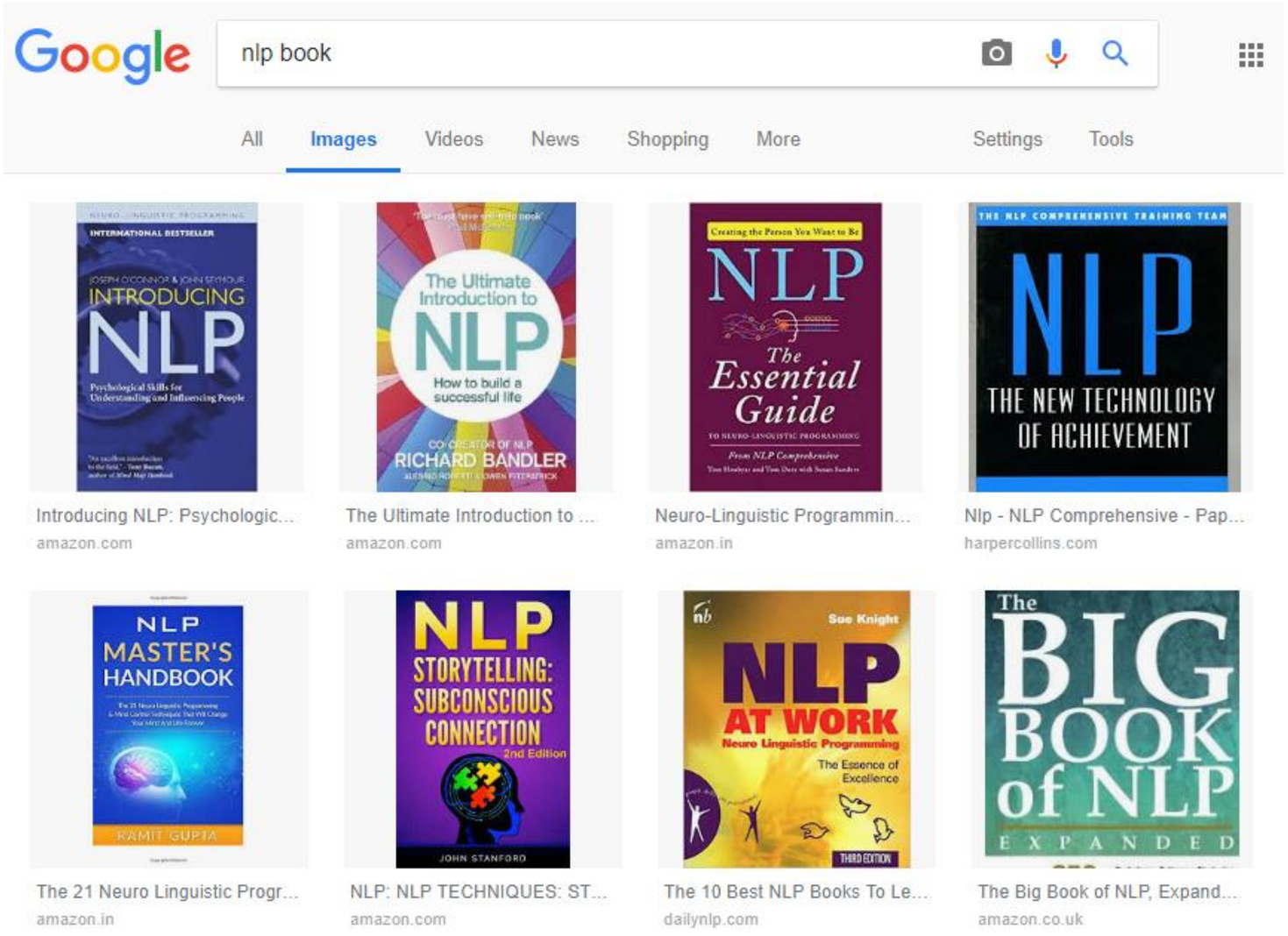


Fig 1. Draw two circles

Fig 2. Draw the rest of the damn owl

What exactly is NLP?

Well… Not this

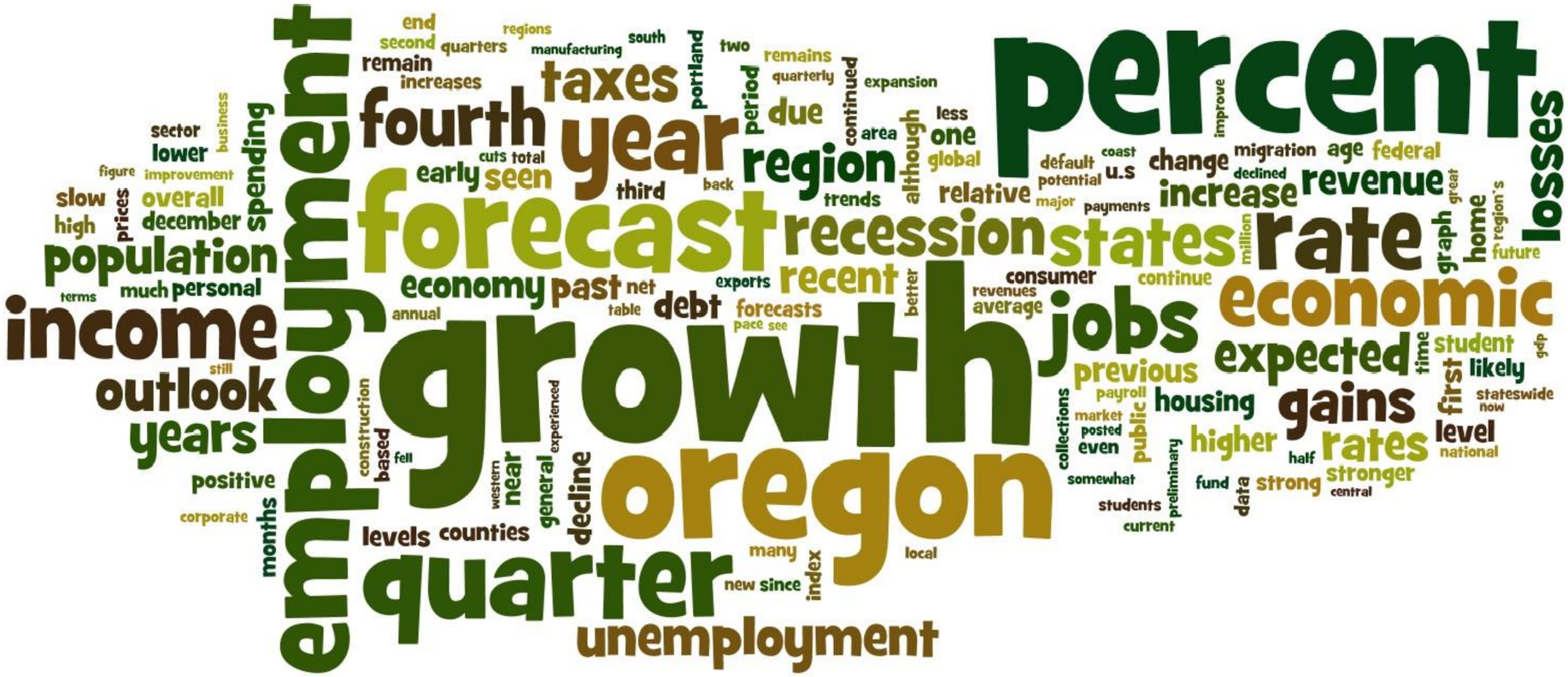


Rather this…


"ALEXA, BUY ME SOMETHING FROM WHOLE FOODS"
amazon
"BUYING WHOLE FOODS"
KAPWING

Natural language processing (NLP) is a subfield of computer science (…) concerned with (…) how to program computers to process (…) natural language data.

How to achieve that?



Source: Oregon Office of Economic Analysis

Tokenization

sentence: "To be or not to be, that is the question."

tokens: | " | To | be | or | not | to | be | , | that | is | the | question |. | " |

Stemming

$$\text{ing} \rightarrow \varepsilon, \quad \text{e.g. running} \rightarrow \text{run}$$

$$\text{ational} \rightarrow \text{ate}, \quad \text{e.g. rational} \rightarrow \text{rate}$$

Sometimes wrong, e.g *booking!*

Stop words


Me
My
We
Myself
It
Our
Hers
He
You
Your
Ours
Yours
She
Her
Him
His
Ourselves
Yourselves
Yourself
They
Who
Their
Has
What
Do
Being
Which
Themselves
Be
Them
Himself
Itself
Had
These
Theirs
Been
Its
Have
Those
Was
Were
Are
Am
Is
This
That
Whom
Having
Herself
Yourself

Bag of Words (BOW)

text 1	*I had a dream*						
text 2	*Dream a little dream of me*						
	i	had	a	dream	little	of	me
BoW 1	1	1	1	1	0	0	0
BoW 2	0	0	1	2	1	1	1

Term frequency – inverse document frequency (tf-idf)

$$tf\text{-}idf(t,d,D) = tf(t,d) \cdot idf(t,D)$$

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$idf(t,D) = ln\left(\frac{|D|}{|\{d \in D : f_{t,d} > 0\}|}\right)$$

$$t = term,\ d = document,\ D = set\ of\ documents,\ f_{t,d} = count(t,d)$$

N-grams

“What a wonderful world”

| <S> What | What a | a wonderful | wonderful world | world <E>

Software packages - code

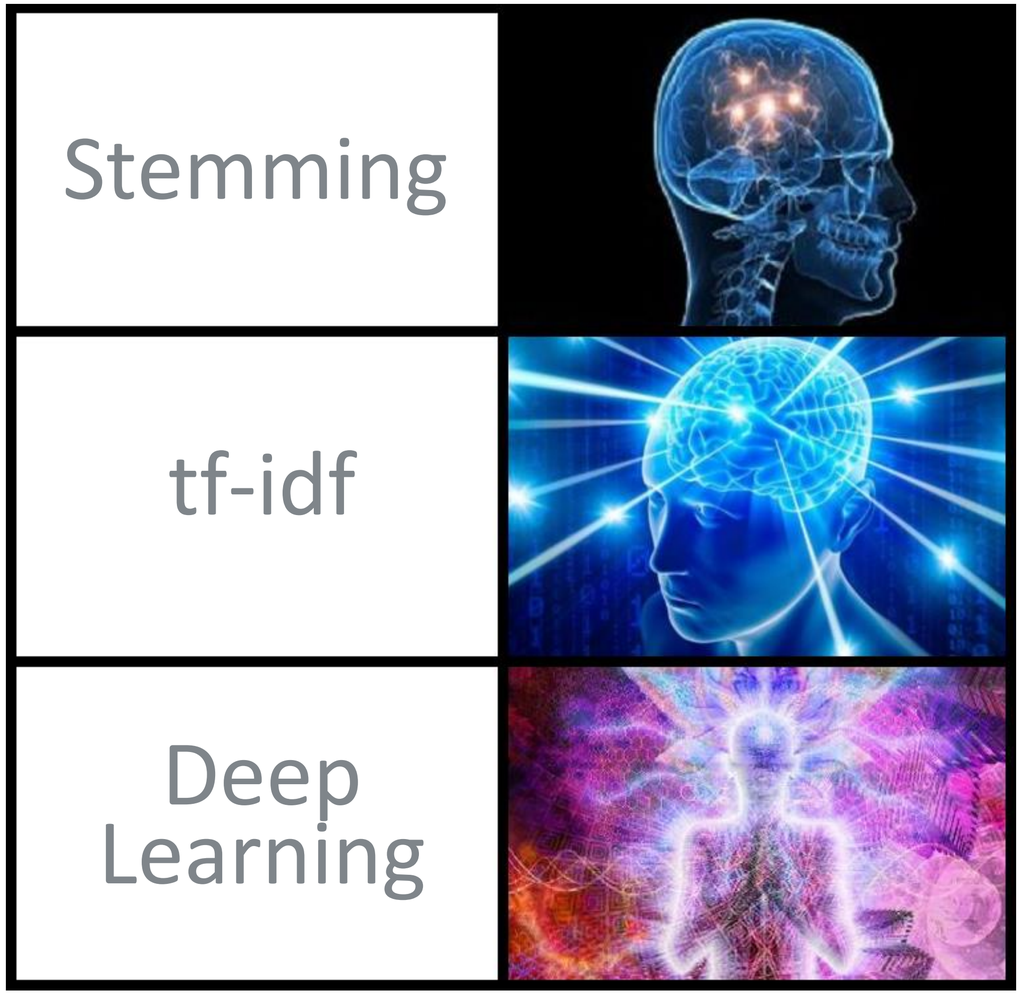
NLTK

spaCy

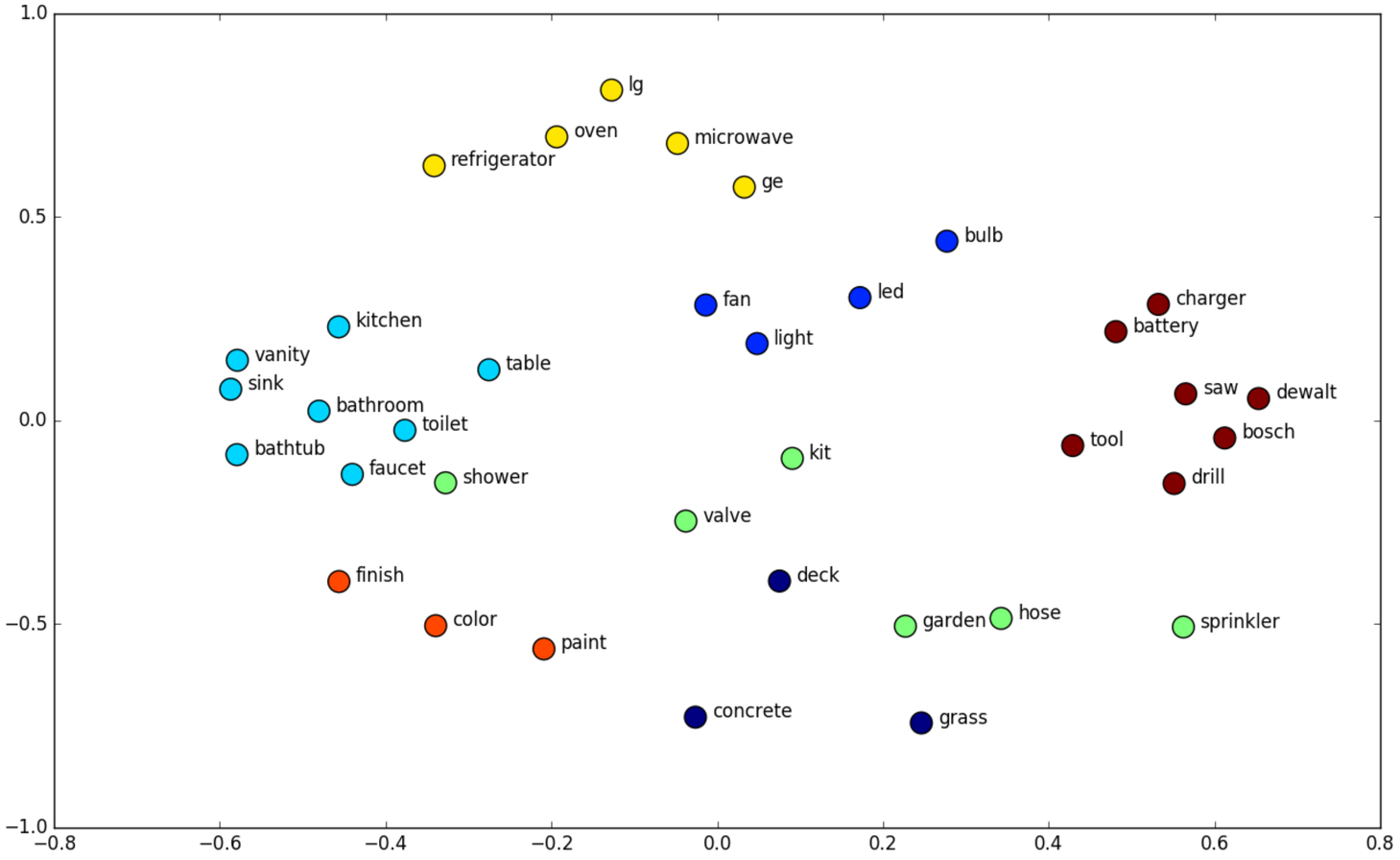
gensim

NLP as a service

Advanced topics



Word Embeddings



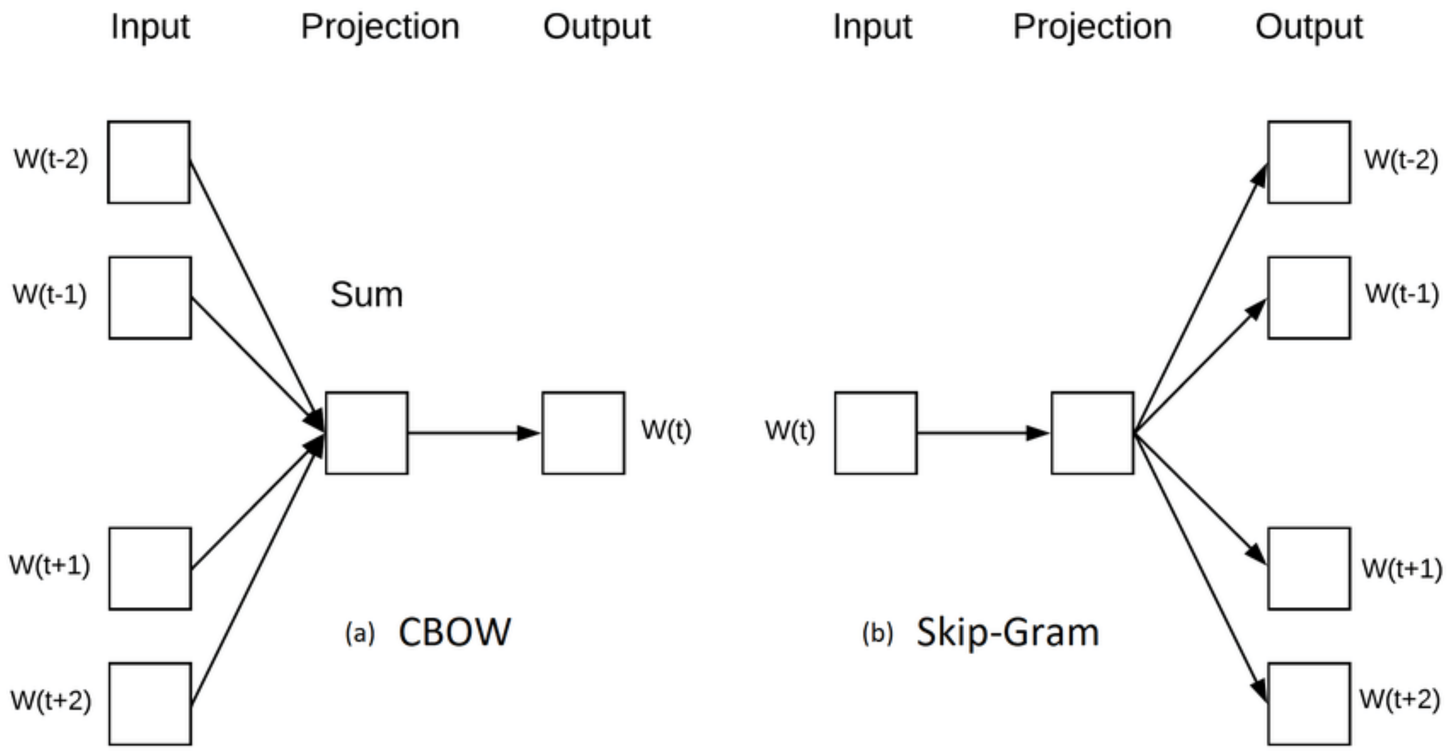
Source: https://www.shanelynn.ie/get-busy-with-word-embeddings-introduction/

„You shall know a word by the company it keeps”

John Rupert Firth

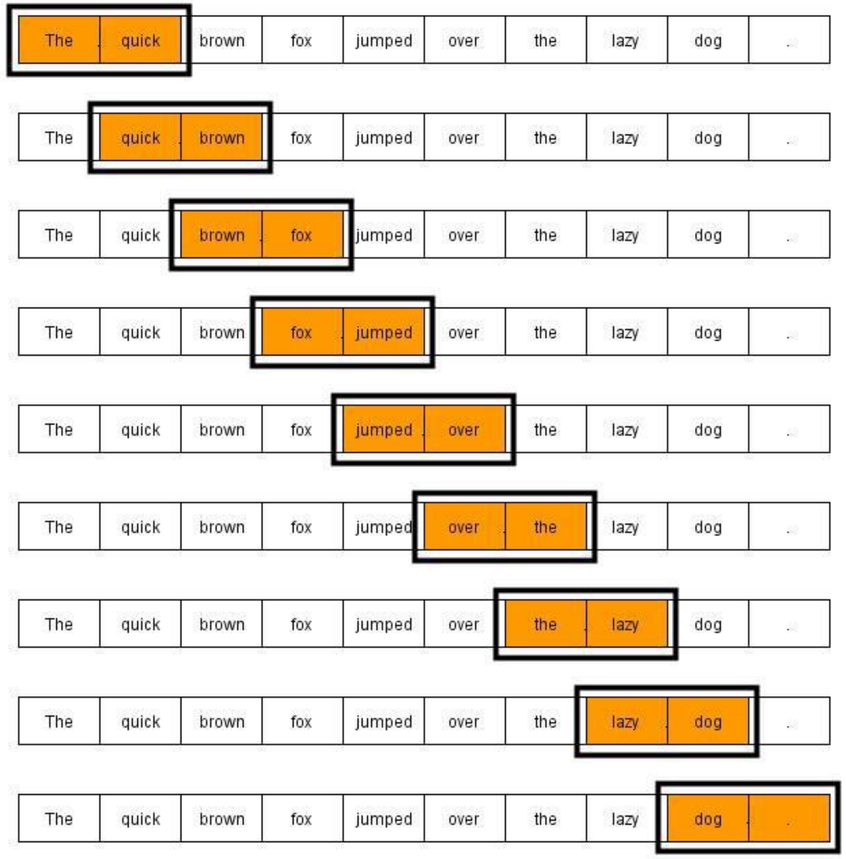
Source: https://cambridge.org

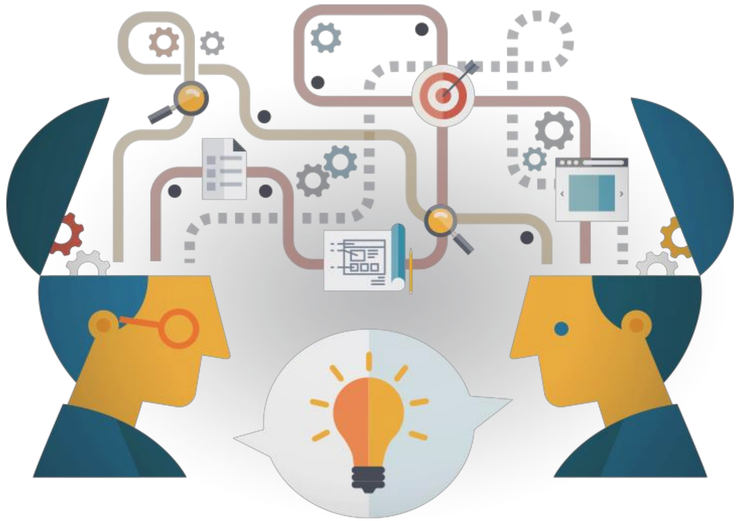
Word Embeddings



Source: https://researchgate.org

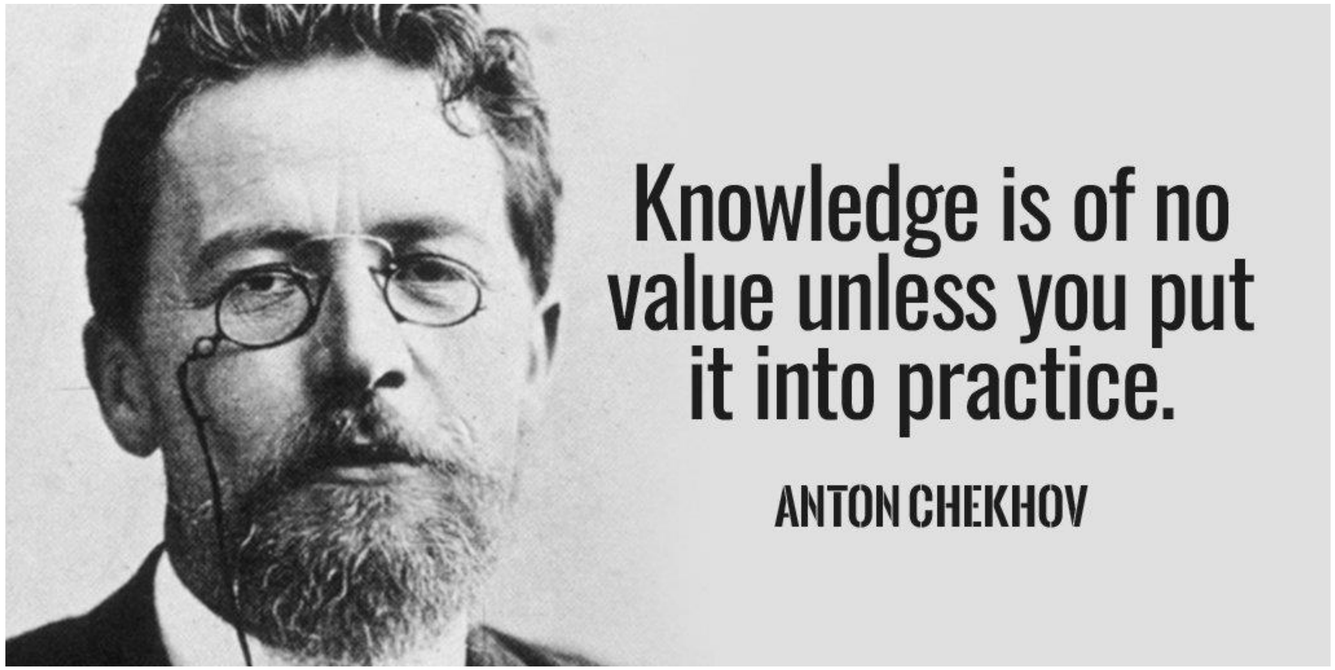
Transfer learning





Source: https://towardsdatascience.com and https://medium.com/paper-club

DEMO


Knowledge is of no value unless you put it into practice.
ANTON CHEKHOV

Thank you